W292–W296 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm344

BIPASS: BioInformatics Pipeline Alternative Splicing Services

Zoé Lacroix[1,*], Christophe Legendre[1], Louiqa Raschid[2] and Ben Snyder[2]

[1]Scientific Data Management Laboratory, Arizona State University, PO Box 875706, Tempe AZ 85287-5706 and [2]Department of Computer Science, University of Maryland, College Park, MD 20742, USA

Received January 31, 2007; Revised April 14, 2007; Accepted April 22, 2007

ABSTRACT

BioInformatics Pipeline Alternative Splicing Services (BIPASS) offer support to scientists interested in gathering information related to alternative splicing (AS) events. The service BIPAS–SpliceDB provides access to AS information that has been extracted *a priori* from various public databases and stored in a data warehouse. In contrast, the BIPAS–Align&Splice service allows scientists to submit their own sequences and genome to compute AS analysis results. BIPAS services offer various user-friendly ways to navigate through the results. AS results are organized at different conceptual levels (clusters and sequences), and are displayed in graphs or summarized in tables that can be downloaded in XML or text format. The two BIPAS services SpliceDB and Align&Splice are available online at http://bip.umiacs.umd.edu:8080/.

INTRODUCTION

Alternative splicing (AS) is the splicing process of a pre-mRNA transcription from one gene that can lead to different mature mRNA molecules, and thus to different proteins. AS has emerged as a major research mechanism after the high-throughput genome sequencing of the 90's and the success of tools that perform pairwise alignment of genomic and transcript sequences. Recent improvements and better accuracy of these alignment tools have demonstrated that the previous assumption of a one-to-one mapping from a gene to a protein no longer holds. Instead AS has come to be accepted as a common process to generate multiple proteins.

AS analysis becomes a critical method for a variety of studies and experiments. AS information may contribute to the reconstitution of events of creation of peptides/proteins from a gene, the discovery of new exons, and thus new proteins, and the understanding of which mechanisms are involved in AS and how they are triggered and regulated. The latter addresses issues related to how cell or tissue characteristics may affect gene translation leading to more than one protein, e.g. a comparison (by computation) of the expression of transcription factor(s) across multiple tissues [see details in (1)]. Additional researches include the discovery of proteins involved in the AS process (splicesosome), the identification of specific sites (such as ESE, ESS, ISE or ISS) involved in the mechanism of AS, the study of activation and repression mechanisms, and some external factors that can influence the AS mechanism, and the identification of possible 'defaults' or 'errors' in AS events that may be directly or indirectly responsible for some diseases [see (2) for more details]. The phylogeny (3) and evolution of AS mechanisms and the identification of specific primers for one specific transcript have also been studied.

FEATURES AND FUNCTIONALITIES

Developed jointly by Arizona State University and the University of Maryland at College Park, the BIPAS services are the first of a suite of BioInformatics Pipeline (BIP) services (4) designed to exploit scientific workflows and database mediation technology to implement scientific pipelines, and to develop useful tools for the AS community.

Technology

We chose IBM's WebSphere Information Integrator (WSII) as the mediator platform (5). Based on the relational data model, WSII makes an autonomous federation of data from heterogeneous sources which appear to the user or front-end application like a single large relational database management system (RDBMS). The SQL query language is supported over all the federated sources even if the underlying sources' native search capabilities are less full-featured than SQL. Similarly, specialized non-SQL search capabilities of the underlying sources are also available through WSII. BIPASS use custom PERL scripts, C++ scripts and

*To whom correspondence should be addressed. Tel: (480) 727 6935; Fax: (480) 965-8325; Email: Zoe.Lacroix@asu.edu

The authors wish it to be known that, in their opinion, all authors should be regarded as joint First Authors.

two alignment tools (Blat and SIM4). The BIPASS server is a DELL PC with two dual-core 64 bit processors operating on RedHat Enterprise Linux 4, 64 bit version. Wrappers were developed to access and retrieve data from multiple public resources, e.g. GenBank (Entrez Nucleotide). BIPAS-SpliceDB exploits data stored *a priori* in the WSII data warehouse, whereas BIPAS-Align-&-Splice processes the users' input and stores the data in the WSII database where the pipeline (alignment and splicing) steps are performed.

Tools

AS analysis is performed in two successive steps. The first is an alignment of a transcript sequence against a genomic sequence, followed by a clustering step. BIPAS services exploit the BioInformatics Pipeline (BIP) toolbox (4).

BIP-Align, the first tool used in the BIP toolbox, can be used alone to create an alignment database, or with other BIP tools to create a database for a more specific function. The objective of BIP-Align is to map input transcripts to a given organism's genome, then store information in a database. Data are subject to user-definable quality filters. The BIP-Align tool extracts and integrates information from several data sources and loads input data in the BIP-SpliceDB, aligns all input transcripts to the genome, and filters and stores alignment data in the BIP-SpliceDB. The current design uses a two-step alignment process that utilizes Blat, then feeds the output through SIM4 to further refine the alignment.

BIP-Splice takes the transcripts loaded and aligned with BIP-Align, clusters them, and performs alternative splicing analysis. The clustering algorithm [used in (1)] is a two-step process. First, transcripts are grouped with respect to overlap, i.e., all transcripts that overlap with at least one base pair are considered part of the same cluster. Overlap takes into account not only the genomic coordinates of the transcripts, but also the orientation, or strand. For instance, if an overlap of two transcripts is based on a genomic position, but one maps to the positive strand and the other to the negative, they are members of two different clusters. The second clustering step uses each transcript's exon/intron structure to refine the clusters. To be a member of a cluster, a transcript must have at least one exon that overlaps with a minimum of one base pair with the exon of another transcript. If it does not meet these criteria then the transcript in question creates a new cluster. The quality filter provides a parameter which requires that all clusters have a minimum number of transcripts as members. The default value is 3, though a BIP-Splice database can be created which allows singleton and doubleton clusters if desired. Each cluster is analyzed to determine whether they exhibit any alternative splicing. The alternative splicing events that are recorded are:

- *Length variation*. This refers to internal splice sites of exons, and if they differ between member transcripts. The 5′ and 3′ ends of the transcripts are not evaluated for splice variation because the sequence may be truncated.
- *Initial cassette exons*. This type of exon is missing in one or more transcripts. An initial exon is the 5′ exon of a transcript. To be flagged as an initial cassette exon, the exon cannot occur as an internal exon in any transcript.
- *Terminal cassette exons*. Same as initial cassette exon, except it occurs at the 3′ end.
- *Internal cassette exons*. These are cassette exons present as internal exons in at least one transcript of the cluster. Internal cassette exons are assumed to be the most biologically relevant because truncated sequences may create artificial occurrences of initial and terminal cassette exons.

Any cluster with at least one form of splice variation is flagged as 'variant'. A cluster typically represents intermediate transcripts [from the pre-messenger-RNA(s) to the mature messenger-RNA(s)] required to obtain one or several functional translated proteins from the same gene. The quality of the alternative splicing analysis depends on several parameters, including the size of the clusters. The more transcripts a cluster contains, The more transcripts a cluster contains, the better the result.

Databases

BIP–SpliceDB creates a data warehouse of data that are automatically extracted using wrappers. Data are obtained from multiple public repositories including UCSC (genome data), GenBank/Entrez Nucleotide (full-length mRNAs) and dbEST (EST data). These databases are a collection of transcript sequences such as cDNA, mRNA, EST, which are often annotated. Annotations lead to a better characterization of the transcripts in each cluster, resulting in an improved accuracy of AS results.

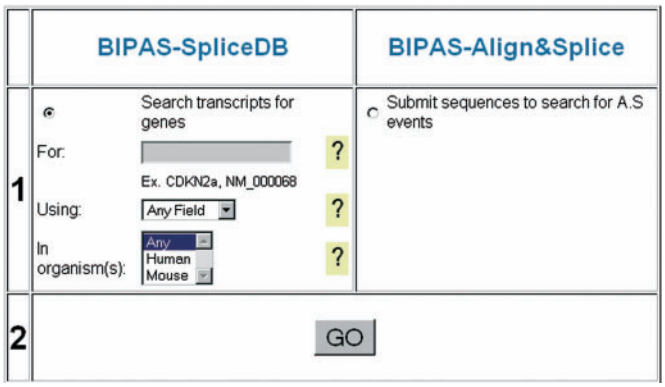


Figure 1. Homepage BIPAS services.

BIPASS USER INTERFACE

The BIPASS front page available at http://bip.umiacs.umd.edu:8080/displays a form (shown in Figure 1) that gives access to two services: BIPAS-SpliceDB on the left and BIPAS-Align&Splice on the right. The display of the site is optimized for Mozilla Firefox 1.5.X.X and 2.0.X.X.

BIPAS–SpliceDB

The first service allows queries against our AS data warehouse.

BIPAS–SpliceDB input

To query BIPAS–SpliceDB, a user enters a keyword and selects the keyword type (Any, Genbank, Annotation) and an organism (Any, Human, or Mouse). When the mouse pad is over one interrogation point some tooltips guide the users. For example, the user enters **cdkn1**, as an annotation, and selects the human genome. Once the query is entered (see Figure 2), the user clicks on the GO button (row 2 in Figure 1) and the results are displayed (see Figure 3).

BIPAS–SpliceDB output

BIPAS-SpliceDB returns a table (Figure 3). The first row of the table indicates the total number of clusters found in the database (this entry has nine clusters). The second row lists the following information describing each cluster (number of clones, clone identifiers, number of genomic exons and whether it contains variants), and information related to the chromosome (identifier, orientation, beginning and end, organism and its genome version). The following rows display information for each cluster matching the user's query.

The user can then click on the cluster of interest; the red arrow in Figure 3 indicates that the user selected the cluster **Hs.chr6.p.17383**. This selection opens a new window showing more details about the cluster including its transcripts, exons, introns, associated splice graph, etc. The details are shown in Figure 4.

The cluster page may be divided in two components: the table and the graph which gives information about the transcripts in the cluster. More links are available at this level. In particular, details on the sequence of genomic exons can be displayed by clicking on the link Sequence(s) in the column 'View'. Links to download data information exist in three formats: XML or text format for the data and png format for the images. In the clickable graph, a click on an exon or intron points directly to its nucleotide sequence in a page containing all the exonic or intronic sequences for the genomic data. By clicking on the label of a transcript a new transcript page containing all information about the transcript is displayed. It contains annotations, information about exons and introns, and a graphical representation of the transcript. At the top of the page, links to download transcript information exist in the three pre-cited formats.

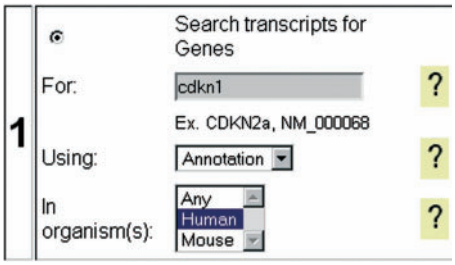


Figure 2. BIPAS–SpliceDB input.

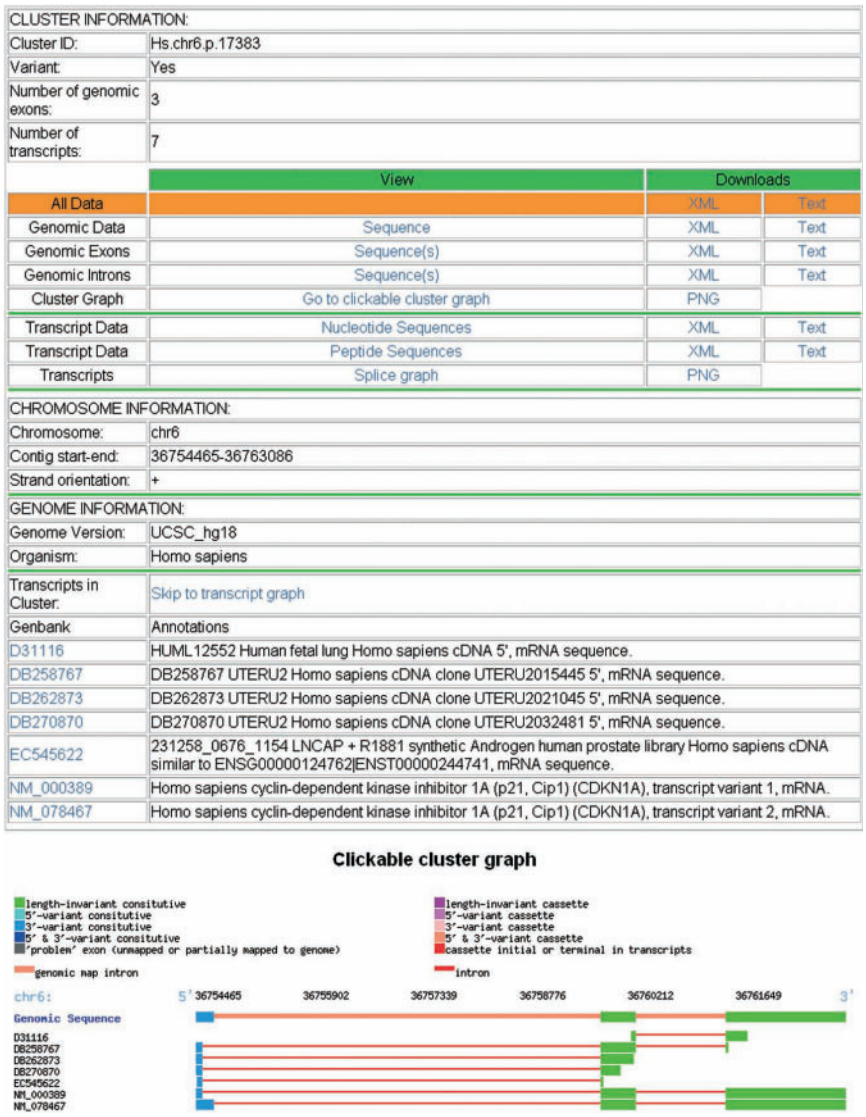
BIPAS–Align&Splice

The second service does not access a data warehouse, in contrast it runs the pipeline online.

Search results:
For: cdkn1
Using: annotation
In organism(s): human

9 clusters found	1 seconds	0	0	Chromosome					
Cluster ID	Number of Clones	Number of Genomic Exons	Variant	ID	Orientation	Begin	End	Genome Version	Organism
1 Hs.chr3.p.2279	3	14	Yes	chr3	+	197951125	198035262	UCSC_hg18	Homo sapiens
2 Hs.chr6.p.17383	7	3	Yes	chr6	+	36754465	36763086	UCSC_hg18	Homo sapiens
3 Hs.chr9.n.6130	7	18	Yes	chr9	-	129968165	130006483	UCSC_hg18	Homo sapiens
4 Hs.chr10.p.8500	9	9	Yes	chr10	+	127502105	127532080	UCSC_hg18	Homo sapiens
5 Hs.chr11.n.11802	2	3	No	chr11	-	2861390	2863579	UCSC_hg18	Homo sapiens
6 Hs.chr15.p.15271	7	12	Yes	chr15	+	38318584	38356979	UCSC_hg18	Homo sapiens
7 Hs.chr19.p.5724	10	11	Yes	chr19	+	44308256	44361886	UCSC_hg18	Homo sapiens
8 Hs.chr20.n.13933	5	12	Yes	chr20	-	9466037	9767680	UCSC_hg18	Homo sapiens
9 Hs.chrX.p.11940	6	23	Yes	chrX	+	110074320	110350829	UCSC_hg18	Homo sapiens

Figure 3. Page of clusters results.



CLUSTER INFORMATION:			
Cluster ID:	Hs.chr6.p.17383		
Variant:	Yes		
Number of genomic exons:	3		
Number of transcripts:	7		
	View	Downloads	
All Data		XML	Text
Genomic Data	Sequence	XML	Text
Genomic Exons	Sequence(s)	XML	Text
Genomic Introns	Sequence(s)	XML	Text
Cluster Graph	Go to clickable cluster graph	PNG	
Transcript Data	Nucleotide Sequences	XML	Text
Transcript Data	Peptide Sequences	XML	Text
Transcripts	Splice graph	PNG	

CHROMOSOME INFORMATION:	
Chromosome:	chr6
Contig start-end:	36754465-36763086
Strand orientation:	+
GENOME INFORMATION:	
Genome Version:	UCSC_hg18
Organism:	Homo sapiens
Transcripts in Cluster:	Skip to transcript graph
Genbank	Annotations
D31116	HUML12552 Human fetal lung Homo sapiens cDNA 5', mRNA sequence.
DB258767	DB258767 UTERU2 Homo sapiens cDNA clone UTERU2015445 5', mRNA sequence.
DB262873	DB262873 UTERU2 Homo sapiens cDNA clone UTERU2021045 5', mRNA sequence.
DB270870	DB270870 UTERU2 Homo sapiens cDNA clone UTERU2032481 5', mRNA sequence.
EC545622	231258_0676_1154 LNCAP + R1881 synthetic Androgen human prostate library Homo sapiens cDNA similar to ENSG00000124762|ENST00000244741, mRNA sequence.
NM_000389	Homo sapiens cyclin-dependent kinase inhibitor 1A (p21, Cip1) (CDKN1A), transcript variant 1, mRNA.
NM_078467	Homo sapiens cyclin-dependent kinase inhibitor 1A (p21, Cip1) (CDKN1A), transcript variant 2, mRNA.

Figure 4. Cluster information page.

BIPAS–Align&Splice input

Once the BIPAS–Align&Splice service is selected, the form shown in Figure 5 is displayed. The completion of the submission is a three-step process:

(1) Enter transcript sequence.
(2) Select or enter genomic data.
(3) Enter a valid e-mail address.

There are two different ways to submit transcript sequences (Paste sequence or Upload sequence from a file) and two different formats that can be used (FASTA or Genbank Format). Several transcript sequences may be submitted at the same time, as long as they are from the same organism and properly formatted. Note that the input must contain at least two exons to return results as the protocol is run online.

The second step is the selection of genomic data. The whole genome of an organism is the default option and the user can decide among one of the available organisms. If the user decides to align ones own transcripts against specific genomic data, the user may enter genomic data as a full consecutive genomic sequence in FASTA format either by uploading or by pasting sequence. For example, one may submit the whole chromosome sequence as long as it is a single sequence in FASTA format.

BIPAS–Align&Splice output

BIPAS–Align&Splice returns a clusters page similar to Figure 3.

BIPASS and AS services

BIPASS are scientist-friendly services dedicated to alternative splicing. BIPAS provides two services directly accessible from its homepage. The BIPAS–SpliceDB service is similar to those provided by Hollywood (6), ASD (7) and H-DBAS (8). BIPAS–SpliceDB is based on

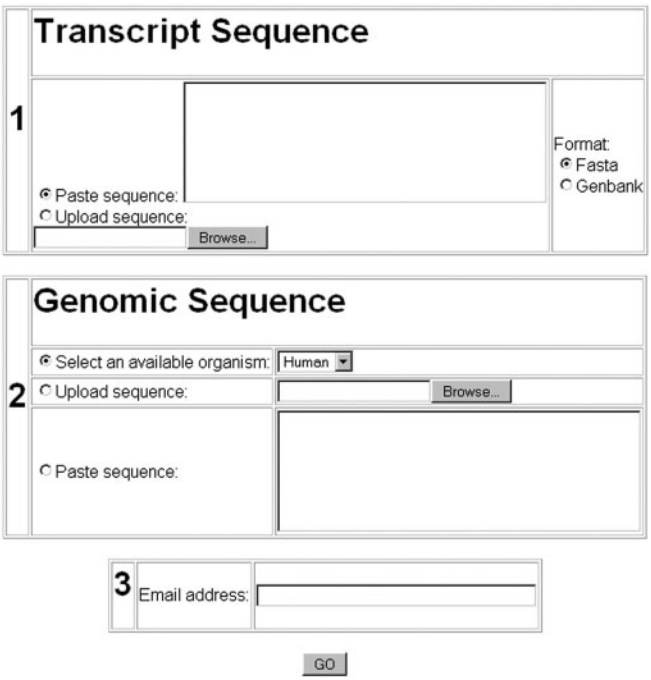


Figure 5. BIPAS–Align&Splice input.

an automatic *a priori* computation of data from different sources both for transcripts and for genomic data. Hollywood, H-DBAS and ASD are also based on the automatic computation of genomic and transcripts data combined with manual data curation. Although manual curation may increase accuracy, it is time and effort consuming. In contrast, automated extraction and computation allows efficient integration of new organisms and data sources. The BIPAS–Align&Splice service is similar to the service provided by ASPIC (9). BIPASS offer a new feature that can align transcripts to a whole genome, allowing users to provide (in one submission) multiple transcripts that are possibly widespread on the genome.

CONCLUSION

BIPAS is an alternative to other AS services. It consolidates two services through a single convenient interface. The first allows the search of a pre-computed BIPAS–SpliceDB warehouse and the second allows the user to analyze AS events on their own transcript sequences. BIPASS will be maintained and improved by updating BIPAS–SpliceDB with new genomes (e.g., rat), providing advanced search features with new search fields (e.g. gene name, sequence) and creating an index of clusters by chromosome.

ACKNOWLEDGEMENTS

This research was partially supported by the National Science Foundation grants IIS0222847, IIS0430915, IIS 0223042, and IIS 0222847. We thank Dr Terry Gaasterland and Dr Bahar Taneri for their contribution to the BIPAS project. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Funding to pay the Open Access publication charges for this article was provided by the National Science Foundation.

REFERENCES

1. Taneri,B., Snyder,B., Novoradovsky,A. and Gaasterland,T. (2004) Alternative splicing of mouse transcription factors affects their DNA-binding domain architecture and is tissue specific. *Genome Biol.*, **5**, R75.
2. Blencowe,B.J. (2006) Alternative splicing: new insights from global analyses. *Cell*, **126**, 37–47.
3. Riegert,P., Wanner,V. and Bahram,S. (1998) Genomics, isoforms, expression, and phylogeny of the MHC class I-related MR1 gene. *J. Immunol.*, **161**, 40667–4077.
4. Eckman,A.B., Gaasterland,T., Lacroix,Z., Raschid,L., Snyder,B. and Vidal,M.E. (2006) Implementing a Bioinformatics Pipeline (BIP) on a mediator platform: Comparing cost and quality of alternate choices. *Proceedings of the 22nd international conference on Data Engneering Workshops*, IEEE 'Computer Society' and Press: Los Alamitos, CA, USA. p. 67.
5. Haas,L., Eckman,B.A., Kodali,P., Lin,E., Rice,J. and Schwarz,P.M. (2003) In Lacroix,Z. and Critchlow,T. (eds), *Bioinformatics: Managing Scientific Data*, Elsevier Science ed. Morgan Kaufmann Publishers, San Francisco, pp. 303.
6. Holste,D., Huo,G., Tung,V. and Burge,C.B. (2006) HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Res.*, **34**, D56–D62.
7. Stamm,S., Riethoven,J.J., Le Texier,V., Gopalakrishnan,C., Kumanduri,V., Tang,Y., Barbosa-Morais,N.L. and Thanaraj,T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46–D55.
8. Takeda,J., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, **35**, D104–D109.
9. Bonizzoni,P., Rizzi,R. and Pesole,G. (2005) ASPIC: a novel method to predict the exon-intron structure of a gene that is optimally compatible to a set of transcript sequences. *BMC Bioinformatics*, **6**, 244.